DOCUMENT RESUME

ED 193 257 TM 800 571

TITLE Standards and Criteria for the Selection of Educational Tests.
INSTITUTION Illinois State Board of Education, Springfield.
PUB DATE [79]
NOTE 16p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Check Lists; Educational Testing; Elementary Secondary Education; *Evaluation Criteria; Test Interpretation; Test Reliability; *Test Selection; Test Validity

ABSTRACT

This manual sets forth comprehension guidelines and procedures for selecting instruments to measure educational growth, achievement, and outcomes. The necessary criteria are presented in a checklist format, permitting local district personnel to evaluate the worth of testing instruments before adoption. A subset of the criteria are defined as "Essential" characteristics, that is, those absolutely necessary for a test to be considered acceptable. The criteria are divided as follows: (1) standards for Manuals, Tests, and Reports; (2) Standards for Validity; (3) Standards for Reliability; and (4) Standards for the Use of Tests. The manual also includes definitions of terms used in conjunction with testing, and of the types of tests (norm referenced, criterion referenced, objective referenced, and domain referenced). (Author/GK)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED193257

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

# Standards And Criteria For The Selection Of Educational Tests

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
C. Reisinger
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

PREPARED BY THE

DEPARTMENT OF PLANNING, RESEARCH, AND EVALUATION
ILLINOIS STATE BOARD OF EDUCATION

TM 800571

## FOREWORD

In 1974, three nationwide professional organizations (American Psychological Association, American Educational Research Association, and the National Council for Measurement in Education) published Standards for Educational and Psychological Tests. This manual sets forth comprehensive guidelines and procedures for selecting instruments to measure educational growth, achievement, and outcomes. A certain subset of the criteria included in this manual were defined as "Essential" characteristics, or characteristics absolutely necessary for a test to be considered acceptable. Due to the current focus upon testing issues in the State of Illinois, it would appear extremely useful for local district personnel to have access to a set of these essential characteristics. This document contains the necessary criteria in a declarative sentence checklist format so that local district personnel can adjudge the worth and strength of testing instruments they are considering adopting.

Joseph M. Cronin
Joseph M. Cronin
State Superintendent of Education



SEP 2 1980

## Introduction and Definition of Terms

Prior to the presentation of the checklist, it would be worthwhile to define certain terms used in conjunction with the testing itself. A test can be defined as any set of items, questions, or tasks which are presented to students to judge educational growth, achievement, or aptitude. This judgement can be made from the perspective of the individual student (i.e., Does the student meet the criterion? Has the student grown since last year?) or from the perspective of the institution (i.e., How well do several of the schools in the district achieve as against national norms? How well does our district compare to other districts in the state?). Regardless of the perspective, the judgments that are made must be reliable and valid. That is, that the scores obtained by students on the test must be consistent each time the students are tested (reliability) and the test items actually do measure that which they are intended to (validity). Later, the various kinds of validity will be pointed out in the checklist.

## Classification of Educational Tests

There are different ways of classifying an educational test. This classification depends upon two characteristics of the test itself. The first characteristic is whether the test uses a known standard of performance with which to judge the students taking the test. Their standards of performance can be either a "norm" or a "criterion." A norm refers to the performance or a test of a particular group of students that can be used for comparisons with a student's achievement. An example of a norm would be:

> "The average score of ninth grade students on this mathematics subtest is 83%."

A criterion refers to a standard established by district personnel, curricular experts, etc. which defines what is considered to be successful performance upon a test. An example of a criterion is:

> "Ninth grade students must score 80% or better upon this mathematics subtest."

Thus, any educational test could be classified as having a standard of performance (norm or criterion) or not.

The second characteristic by which educational tests are classified is in relation to whether the items of a test are keyed to a particular and fixed set of curricular objectives. An objective is any specific statement as to what skills or knowledge a student should attain due to the education that student receives. An example of an objective is:

> "Ninth grade students in mathematics should be able to compute with rational numbers."

Thus, any educational test can be classified in terms of whether its items are derived from or keyed to a set of curricular objectives.

These two characteristics of tests (standards of performance and particular objectives) are extremely important in regard to determining the nature and purpose of an educational assessment instrument. In fact, the definition of what type of test a measurement instrument is (criterion referenced, norm referenced, objective referenced, domain referenced) depends upon its classification with those two characteristics. The typology of tests with regard to these characteristics is displayed in Figure 1.

Criterion Referenced Tests

From Figure 1 it can be seen that a criterion referenced test has items which are uniquely keyed to a specific set of objectives and curriculum. Moreover, a criterion referenced test has a very definite standard or criteria for successful performance.

Criterion referenced tests may be used for the purpose of addressing issues of "pupil promotion" and individual student achievement. Criterion referenced tests can also be used to measure an individual student's growth across time against certain curricular goals.

Norm Referenced Tests

Norm referenced tests, like criterion referenced tests also contain standards of performance. For norm referenced tests this standard is the "norm" group against which the individual student scores are compared. However, unlike criterion referenced tests, norm referenced tests contain items which are not tied to particular objectives. The items may be pooled into subtests (mathematics, reading, etc.) but are not individually keyed to specific objectives.

The optimal use of norm referenced tests is both for pupil "diagnosis" and pupil "growth." In terms of diagnosis, norm referenced tests will be used to assess a student's cognitive strengths and weaknesses when compared to the overall population of similar students. Likewise, norm referenced tests may be used to measure growth and development of a student. However, the standard of comparison for norm referenced tests is to judge whether a student is growing and changing like students of his/her same age are growing and changing. (This is a contrast to criterion referenced tests which judge growth against a student's own base line.)

Objective Referenced Tests

Objective referenced tests, like criterion referenced ones do contain items which are specifically keyed to a set of objectives. However, objective referenced tests do not have any standard, criteria, or norm of successful performance. Thus, objective referenced tests are uniquely fitted for the purpose of "pupil placement." Since its items are keyed to local district curriculum, a student's scores upon the test may indicate where that student can best profit from the various instructional programs and curriculum levels offered by a particular school district.

Moreover, objective performance tests can be used to obtain an overall perspective on how well the district as a whole, or subunits inside the district (schools etc.) are scoring. A district could evaluate its own strengths and weaknesses (or strengths and weaknesses of particular schools) in terms of its own objectives through the use of one of these tests. (Note that the Illinois Inventory of Educational Progress is an objective referenced test.)

Domain Referenced Test

Domain referenced tests are those which have neither standards of performance nor items keyed to particular objectives (see Figure 1). They should be seen as a pool of items tapping a very general domain, such as mathematics. Their purpose is simply to estimate how well a student will perform in the future over a large number of similar items. If a district can define its goals or curriculum quite globally or generally, then a domain referenced test could be used for diagnostic purposes. Domain referenced tests have also been used to measure the impact of new, innovative, compensatory, or experimental courses or training upon a student's subsequent achievement.

Concluding Caveat

It must be remembered, however, that testing provides only one indicator of educational growth and achievement and that multiple perspectives should be brought to bear in evaluating students.

With these definitions, we can proceed to present the checklist of test characteristics.

| OBJECTIVES | STANDARDS OF PERFORMANCE: Present | STANDARDS OF PERFORMANCE: Absent |
|---|---|---|
| Present | CRITERION Referenced | OBJECTIVES Referenced |
| Absent | NORMATIVE Referenced | DOMAIN Referenced |

Classification of Educational Tests

(FYANS, 1978)

Figure 1

## I. Essential Standards for Manuals, Tests, and Reports

| | YES | NO |
|---|---|---|
| **(A) Dissemination of Information** | | |
| (1) The test is accompanied by a manual of information from the test publisher. | ___ | ___ |
| (2) The manual comprehensively defines the process of development of the test (rationale, item development, analysis). | ___ | ___ |
| (3) All revised versions of a test should have their own manual (or revised original manuals). | ___ | ___ |
| **(B) Aid to Interpretation** | | |
| (1) The manual clearly describes influences on test performance associated with socio-economic status, sex, ethnicity, race, or creed. | ___ | ___ |
| (2) The test manual states explicity the applications and purposes of the test. | ___ | ___ |
| (3) The test manual describes the psychological or educational theory underlying the test. | ___ | ___ |
| (4) For criterion-referenced tests, the content and objectives of content domain are well-defined. | ___ | ___ |
| (5) For mastery tests, the cutting score is suggested or methods by which users could form cutting scores are defined. | ___ | ___ |
| (6) The manual does not imply that the test is "self-interpreting," but suggests interpretation strategies. | ___ | ___ |
| (7) The statistical report in the manual is one that clearly supports the statements made concerning the test. | ___ | ___ |

| | YES | NO |
|---|---|---|
| (8) If computer based interpretations of scores are available, their rationale and evidence of their interpetative capability is provided. | ___ | ___ |
| **(C) Administration Directions** | | |
| (1) The directions are clear enough for the test user to duplicate the conditions of the norming of the test. (e.g. is guessing, allowed, time limits, marking answer sheets) | ___ | ___ |
| (2) The scoring of the test is clearly understandable. | ___ | ___ |
| **(D) Norms and Norming Processes** | | |
| (1) Norms are available in the manual when the test is released. | ___ | ___ |
| (2) The norms provided in the manual are for clearly defined populations and those which the user will wish to use for comparisons. | ___ | ___ |
| (3) The sampling method followed well-defined procedures and was not based upon convenience or readily available populations. | ___ | ___ |
| (4) If the test will be used for subgroups, their norms on the test are provided. | ___ | ___ |
| (5) The manual clearly indicates the number of students in the norming sample and the number of sampling units (strata, etc.) | ___ | ___ |
| (6) When the test is to be used to assess groups (programs, etc.) of students rather than individuals, normative data for the group statistics are provided. | ___ | ___ |

| | YES | NO |
|---|---|---|
| (7) When a new form of a test is equated with the older form, the revised manual describes the content of both forms and their norm groups. | ___ | ___ |

(E) Scales of Test Scores

| | | |
|---|---|---|
| (1) Any scales of scores (e.g. percentile rank; standard scores; T-scores) derived from the raw scores are clearly described and interpreted in the manual. | ___ | ___ |
| (2) The manual specifies whether derived scores are linear transformations (e.g. standard scores) or normalized (e.g. percentile ranks). | ___ | ___ |

II. Essential Standards for Validity

(A) General Principle of Validity

| | | |
|---|---|---|
| (1) The manual (or reports provided) present evidence of validity for each recommended purpose of the test (no test is valid for all purposes or in all situations). | ___ | ___ |
| (2) If the use of profiles, subtest scores, or score differences are recommended, the evidence justifying the recommendation is given. | ___ | ___ |
| (3) The test manual indicates that correlations between the items and the total test itself are item discrimination not item validity indices. | ___ | ___ |

(B) Criterion Related Predictive Validity

| | | |
|---|---|---|
| (1) The manual (or reports supplied) clearly define the adequacy of any criteria to be predicted by the test score performance. | ___ | ___ |

| | YES | NO |
|---|---|---|
| (2) The evidence for validity for all criteria the test is purported to predict is supplied in the manual (or reports supplied). | ___ | ___ |
| (3) The samples used in the validation process are described such as age, sex, socio-economic status, ethnic origin, residential region, and level of education. | ___ | ___ |
| (4) The test manual (or reports provided) clearly indicates for how long of a time predictions can be made from the test. | ___ | ___ |
| (5) Separate evidence is provided for long term prediction and concurrent prediction. (Evidence for one cannot be used for the other). | ___ | ___ |
| (6) If the manual mentions that the predictive validity coefficients are corrected for errors of measurement or selection both the corrected and uncorrected coefficients are reported. | ___ | ___ |

(C) Content Validity

| | | |
|---|---|---|
| (1) If test performance is to be interpreted as a sample of typical performance of a student in a content area, the test manual clearly defines the content area used in the test. | ___ | ___ |
| (2) Information should be provided as to the date when the items were keyed to a particular content area. | ___ | ___ |

| | YES | NO |
|---|---|---|
| (D) Construct Validity | | |
| (1) If the test is given within a time limit, evidence is provided concerning the effect of speed on scores and their correlation with other variables. | ___ | ___ |
| (2) If the test is purported to measure an "ability" or "aptitude" or "personality trait," those aptitudes, abilities, or traits are clearly defined and how the scores on the test are to be interpreted for that aptitude (etc.) are given. | ___ | ___ |
| III. Essential Standards for Reliability | | |
| (A) General Principles of Reliability | | |
| (1) The test manual (or reports supplied) presents evidence of the reliability of the test including estimates of standard error of measurements. | ___ | ___ |
| (2) The procedures and samples used for establishing the reliability of the test (standard errors of measurement) are clearly described. | ___ | ___ |
| (3) If the manual mentions that the reliability coefficients are corrected for restriction in range, both the corrected and uncorrected coefficients should be presented. | ___ | ___ |
| (4) If the test is to be used for district subgroups (sex, ethnic, age groups, etc.) separate reliability information should be given for each group. | ___ | ___ |
| (B) Comparability of Test Forms | | |
| (1) If two or more forms of a test are provided, their respective means, standard deviations; item characteristics, and inter-correlation are provided to establish their equivalence. | ___ | ___ |

| | YES | NO |
|---|---|---|
| (C) Internal Consistency of a Test | | |
| (1) Estimates of the internal consistency of the test is given. (Internal consistency information is not required for speeded tests.) | ___ | ___ |
| (D) Comparability Over Time | | |
| (1) The test manual indicates to what extent the test scores are stable and constant across time. | ___ | ___ |
| (2) The manual indicates the effect of using a different form of the test upon the comparability of the scores. | ___ | ___ |
| (3) If the test is to be used to compare groups of students, rather than individuals, standard errors and standard errors of measurement for group means are presented. | ___ | ___ |
| IV. Essential Standards for the Use of Tests | | |
| (A) Concern of Users | | |
| (1) Test users will periodically review the use and purpose of the test. | ___ | ___ |
| (B) Choice of Testing Instrument | | |
| (1) The choice of a test, test battery, or assessment procedure are based on a local education agency's clearly defined goals goals and objectives. | ___ | ___ |
| (2) The test represents balanced coverage of the districts goals. | ___ | ___ |
| (3) Several methods of assessing a students ability or knowledge are considered and used. | ___ | ___ |
| (4) The test is in the dominant language of individuals tested. | ___ | ___ |

| | YES | NO |
|---|---|---|
| **(C) Administration and Scoring** | | |
| (1) The instructions for the test are appropriate for individuals tested. | ___ | ___ |
| (2) The format for the test is appropriate for individuals tested. | ___ | ___ |
| (3) The response form for test is appropriate for individuals tested. | ___ | ___ |
| (4) The test is administered under standardized conditions as specified in the manual. | ___ | ___ |
| (5) Testing is within 2 weeks of mid point norm of dates. | ___ | ___ |
| (6) If test scoring equipment is used, periodic checks are made of its accuracy in scoring, checking, coding, and recording test results. | ___ | ___ |
| (7) The user has not made substantial changes in the test format, context, instructions, language, and content. | ___ | ___ |
| (8) Reasonable precautions are taken to safeguard test security. | ___ | ___ |
| **(D) Interpretation of Test Scores** | | |
| (1) A test score is interpreted as a estimate of performance under a given set of circumstances and is not interpreted as an absolute and permanent characteristic of the student. | ___ | ___ |
| (2) The user recognizes that estimates of reliability do not indicate the predictive validity of the test. | ___ | ___ |

| | YES | NO |
|---|---|---|
| (3) The test user interprets the scores of students in terms of scales such as percentile ranks or standard scores and not less representative terms such as grade-equivalents. | ___ | ___ |
| (4) Any content referenced interpretation clearly states the domains (e.g. mathematics ability, reading ability) to which that interpretation will generalize. | ___ | ___ |
| (5) Any norm referenced interpretation is with reference to the appropriate set of norms for the individuals tested. | ___ | ___ |

## References

Fyans, Leslie, J., Jr. (Instructor) The Generalizability of Educational Measurement. Pre-session Training Courses. American Educational Research Association. Toronto, Canada 1978.

Standards for Educational and Psychological Tests. American Psychological Association. Washington, D.C., 1974.

CMC/1362